ROBUST BAYESIAN ANALYSIS AND OPTIMAL EXPERIMENTAL DESIGNS IN NORMAL LINEAR MODELS WITH MANY PARAMETERS–II

by

Anirban DasGupta * and William J. Studden **

Technical Report #88-34C

Department of Statistics
Purdue University

August 1988

* Research supported by NSF Grant DMS 87026–20
** Research supported by NSF Grant DMS 88025–35

★

Abstract

In a novel approach to experimental design, we address the problem of finding a design that minimizes the Bayes risk with respect to a fixed elicited prior subject to being robust with respect to misspecification of the prior. Uncertainty in the prior is formulated in terms of having a family of priors instead of one single prior. Two different classes of priors are considered: a family of conjugate priors, and a second family of priors induced by a metric on the space of nonnegative measures. Family 1 has earlier been suggested by Leamer (1978, 1982) and Polasek (1984), while family 2 was considered in DeRobertis and Hartigan (1981), and Berger (1987). The setup assumed is that of a canonical normal linear model with independent homoscedastic errors. Optimal designs are worked out for the problem of estimating the vector of regression coefficients or a linear combination of the regression coefficients and also for testing and set estimation problems. Some new convexity results are established and concrete examples are given for polynomial and weighted polynomial regressions and a completely randomized design. A very surprising finding is that for family 2, the same design is optimal for a variety of different problems with different loss structures. In general, the results for family 2 are significantly more substantive. Our results are applicable to group decision making and reconciliation of opinions among experts with different priors.

1. Introduction. A major problem in the general domain of statistics is the derivation of an experimental design optimal with respect to some criterion consistent with the goal of the study. Typically, the optimality criteria considered by workers in this general area have focused on long run (frequentist) performance of a design, such as the mean squared error over repeated sampling: the well known criteria of A, D and E optimality are examples of this kind. It is not unusual though for the experimenter to have nonnegligible prior information about the parameters in the system, information that is sufficiently significant to be of some use but not quite so sharp and precise as to be quantified in terms of a single "prior distribution." The purpose of this article is to address the question of which design should the statistician recommend in the scenario of a collection of plausible, Bayesian prior distributions. This article thus focuses on some experimental design problems from a "robust Bayesian" viewpoint. The subject of robust Bayes methods has, by itself, been a major research area in the recent past; for general exposition and specific results, we refer the reader to Berger (1984, 1987), Berger and Berliner (1984), Leamer (1978, 1982), Polasek (1984), DeRobertis and Hartigan (1981), Kadane and Chuang (1987), Good and Crook (1987), Lindley and Smith (1972), Dempster (1975), West (1979), Duncan and Lambert (1981), and DasGupta and Studden (1988a, 1988b, 1988c, 1988d) etc.

There now exists a vast body of statistical literature on optimal experimental designs (with primarily long run performance criteria); the pioneering work is due to Jack Kiefer. For a variety of results and general exposition, see Kiefer (1959, 1961, 1974), Kiefer and Wolfowitz (1959), Sacks and Ylvisaker (1968), Brooks (1972, 1974), Chernoff (1972), Duncan and DeGroot (1976), Elfving (1952), Fedorov (1972), Hoel (1966), Lindley (1968), Pukelsheim and Titterington (1983), Silvey (1980), Whittle (1973), Cheng (1987), Wynn (1972), Karlin and Studden (1966), etc.

The study of experimental designs in a Bayesian framework has been comparatively limited; some of the important references include Pilz (1979, 1981), Verdinelli (1982), Bandemer (1977), Chaloner (1984) etc. In this article, optimal experimental designs are derived for the problems of estimation, prediction, or testing a null hypothesis in the canonical normal linear model set up when the prior distribution for the parameters belongs to a family of distributions (measures) Γ.

Consider the usual linear regression problem where $\underset{\sim}{Y}_{n\times 1} \sim N(X\underset{\sim}{\theta},\ \sigma^2 I)$, where $X_{n\times p}$ is the design matrix of nonstochastic constants; for ease in exposition, assume $\sigma^2 > 0$ to be known; σ^2 comes out as a proportionality factor in all risk expressions relevant to this paper and consequently will be ignored in all risk formulas. The case of the unknown error variance will be mentioned in the concluding section. The design aspects of the problem enter through the experimenter's choice of the rows of the design matrix X from an available set $\mathcal{X}$. The vector of regression coefficients $\underset{\sim}{\theta}_{p\times 1}$ is assumed to have a prior distribution $\pi(\underset{\sim}{\theta})$ belonging to a suitable class Γ. For example, if Γ is the class of all multivariate normal distributions with a fixed mean $\underset{\sim}{\mu}$ and a variance–covariance matrix $\sigma^2\Sigma$, where $\Sigma_1 \leq \Sigma \leq \Sigma_2$ (in the sense that $\Sigma - \Sigma_1$ and $\Sigma_2 - \Sigma$ are nonnegative definite) for two fixed matrices Σ_1 and Σ_2, then a relevant experimental design problem would be to choose a design matrix X such that the variations in Bayesian measures of interest (such as the Bayes risk or the Bayes estimate itself) due to the uncertainty in the prior is minimized among all possible designs X. If such a design can be worked out, then we can legitimately describe it as the "most robust" design. A moment's reflection shows, however, that designing simply to obtain the most robust results can arguably result in a collection of statistical estimators which are similar in magnitude, but are mostly wrong or have other undesirable properties. It seems natural, therefore, to use robustness as a secondary criterion at the design stage, the primary goal being near Bayesness with respect to a fixed elicited prior. Here is a simple example. In the situation described above, the Bayes risk under ordinary squared error loss $L(\underset{\sim}{\theta}, \underset{\sim}{a}) = ||\underset{\sim}{\theta} - \underset{\sim}{a}||^2$ when the variance–covariance matrix of $\underset{\sim}{\theta}$ equals $\sigma^2\Sigma$ is equal to $tr(X'X + \Sigma^{-1})^{-1}$. The range of the possible Bayes risks is therefore $tr(X'X + \Sigma_2^{-1})^{-1} - tr(X'X + \Sigma_1^{-1})^{-1}$. A sensible formulation of the design problem would then be to minimize $tr(X'X + \Sigma_2^{-1})^{-1} - tr(X'X + \Sigma_1^{-1})^{-1}$ subject to the restriction that $tr(X'X + \Sigma_0^{-1})^{-1} \leq (1+\varepsilon) tr(X_0'X_0 + \Sigma_0^{-1})^{-1}$ where Σ_0 is a fixed matrix ($\Sigma_1 \leq \Sigma_0 \leq \Sigma_2$), X_0 is the Bayes design with respect to Σ_0, and $\varepsilon \geq 0$ is a fixed (usually small) real number. It is also quite natural to try to minimize the variation in the Bayes estimate of $\underset{\sim}{\theta}$ itself. To be more precise, if one wants to estimate the vector $\underset{\sim}{\theta}$ (or predict k future values of the response variable y) under a squared error loss, then the collection of all possible Bayes estimates (or Bayes predictors) as the prior π for $\underset{\sim}{\theta}$ varies in the class Γ will usually form a nice convex set S in an euclidean space. We could define a metric d on the set S and minimize the (expected) diameter $E(D)$ of S for the metric d (note $D = \sup_{\underset{\sim}{u},\underset{\sim}{v} \varepsilon S} d(\underset{\sim}{u}, \underset{\sim}{v})$), again subject to suitable restrictions like $tr(X'X + \Sigma_0^{-1})^{-1} \leq (1+\varepsilon) tr(X_0'X_0 + \Sigma_0^{-1})^{-1}$. The expected value of D in this calculation could be taken under the marginal distribution of the response variable induced by the $N(\underset{\sim}{\mu}, \Sigma_0)$ prior (the reason we have to take an expected value of D is that D itself will be a function of the specific data obtained, but we do not see such data at the design stage). Design problems of these types will be addressed in this article.

For ease of exposition we shall consider, what is now commonly called, the approximate design theory. All the design aspects enter through the "information matrix" $X'X$ which can be written as $X'X = n\Sigma p_i x_i x_i'$ where x_i' are the rows of X and $np_i = n_i$ are integers. The approximate theory allows the $p_i \geq 0$ to be arbitrary, subject to $\Sigma p_i = 1$, and in fact, for further convenience, permits $X'X = n\int xx'd\mu$ where μ is an arbitrary probability measure on $\mathcal{X}$.

Two different classes of priors will be considered; the first of them is

$$\Gamma_1 = \{\pi(\underset{\sim}{\theta}) : \underset{\sim}{\theta} \sim N(\underset{\sim}{\mu}, \sigma^2\Sigma), \underset{\sim}{\mu} \text{ fixed}, \Sigma_1 \leq \Sigma \leq \Sigma_2\}; \tag{1.1}$$

the idea here is that conjugate priors are mathematically attractive and also often provide a rich enough class of priors for an honest Bayesian analysis of the data; the mean of the prior is kept fixed but not the variance–covariance structure because the location of the unknown parameters is usually much easier to elicit subjectively than it is to elicit the higher moments and the strengths of the correlations. Also, as we shall later see, the design problems are reasonably tractable with a family of priors such as (1.1). The family of priors (1.1) was first suggested and used by Leamer (1978, 1982), and Polasek (1984). For an extensive discussion, see DasGupta and Studden (1988a).

Normal priors, by definition, are symmetric and unimodal. Moreover, in (1.1) the mean $\underset{\sim}{\mu}$ was kept fixed (although we could vary the prior mean as well; see DasGupta and Studden (1988a)). An alternative family of priors that also enjoys mathematical tractability, and yet at the same time automatically changes the mean along with the variances and the covariances and in addition includes asymmetric and multimodal priors is the family of priors

$$\Gamma_2 = \{\pi(\underset{\sim}{\theta}) : L(\underset{\sim}{\theta}) \leq \pi(\underset{\sim}{\theta}) \leq U(\underset{\sim}{\theta})\}, \tag{1.2}$$

where L and U are two fixed nonnegative functions, not necessarily probability densities (i.e., L and U may not integrate to 1). Roughly speaking, the class of priors (1.2) places the prior within a fixed band, much as one constructs confidence bands for the response curve in regression problems. In applications, if one takes the lower band L as the density function of a fixed distribution and U as kL where $k > 1$ is a fixed number, then the family (1.2) is a metric neighborhood of the prior L; again, see DasGupta and Studden (1988a). The family (1.2) has many other highly attractive features and is thoroughly discussed in sections 1 and 6 in DasGupta and Studden (1988a). The first works with this family of priors are DeRobertis (1978) and DeRobertis and Hartigan (1981).

Section 2 contains the optimal design results for the family of priors (1.1). A general result on the extremities of a convex functional is also proved in this section. This result, although quite simple, may be of independent interest. In section 3, we derive the optimal designs for the family of priors (1.2). It is seen that the design which is Bayes with respect to the prior L has many robustness properties. For example, the following result is proved: if L is a $N_p(\underset{\sim}{\mu}, \sigma^2\Sigma)$ density and $U = kL$ for some $k > 1$, the regular Bayes design against the prior L minimizes the (expected) Euclidean diameter of the set of Bayes estimates of $\underset{\sim}{\theta}$. A variety of other optimal design results are proved in both sections 2 and 3. The analysis is substantially more elegant for Γ_2. However, the family Γ_1 needed to be considered because of its conventional nature. Mathematically, our article contains new convexity results and counterexamples to demonstrate that some functionals of the information matrix $X'X$ that one would expect to be convex are in fact not necessarily convex. All through the article, the results from DasGupta and Studden (1988a) are heavily borrowed. We hope that our results will be useful in the context of group decision making and reconciliation of opinions in the face of imprecise but significant prior information.

2. Normal priors with a fixed mean. In this section, we consider several optimal design problems when we have the family of priors (1.1). We are interested in minimizing various functionals, for example, $\Phi_1(M) = tr(M+\Sigma_2^{-1})^{-1} - tr(M+\Sigma_1^{-1})^{-1}$ subject to the condition that $\Phi_0(M) = tr(M+\Sigma^{-1})^{-1}$ is near its minimum. Thus, if M_0 minimizes Φ_0, we require the minimum of $\Phi_1(M)$ subject to $\Phi_0(M) \leq (1+\varepsilon)\Phi_0(M_0)$ for some specified $\varepsilon > 0$. Here $\Phi_0(M)$ is proportional to the Bayes risk with prior corresponding to Σ and Φ_1 is proportional to the range of the Bayes risks for π in (1.1). The main difficulty in obtaining useful results to aid in obtaining solutions to such problems is that the matrix difference itself, $(M+\Sigma_2^{-1})^{-1} - (M+\Sigma_1^{-1})^{-1}$ is, in general, neither decreasing nor convex in M. Each of the terms $(M+\Sigma_i^{-1})^{-1}$ is, of course, decreasing and convex in M, so that $\Phi_0(M)$ is decreasing and convex. Counter examples in this regard will be given after example 2 below. Whenever Φ_1 is convex and decreasing one can then appeal to the usual Lagrangian theory for such constrained problems. In these cases the minimum will generally satisfy the equality $\Phi_0(M) = (1+\varepsilon)\Phi_0(M_0)$ provided the global minimum of Φ_1 is not already in the constraining set.

If Φ_0 itself is not convex (or not decreasing, or both), it is still very useful to know that the minimum for the constrained problem will occur when equality is present. This usually provides some reduction in the dimensionality of the problem. We state and prove a general theorem in this regard.

Theorem 2.1. Let Z be a locally convex compact topological vector space and f_1 be a continuous convex function on Z, and f_0 a continuous function on Z, not necessarily convex. Let $S = \{z\epsilon Z: f_0(z) \leq k\}$ where k is such that $\inf_z f_0(z) \leq K < \sup_z f_0(z)$. Suppose there exists a unique $z_1 \epsilon Z$ such that $f_1(z_1) = \inf_z f_1(z)$. Then either $z_1 \varepsilon S$ or the infimum of f_1 on S is attained at a z^* such that $f_0(z^*) = k$.

Remark 1. Note that since f_1 is continuous and Z compact, z_1 always exists; moreover z_1 is unique if f_1 is strictly convex.

Proof of Theorem 2.1: Suppose $z_1 \notin S$. Note S is closed and hence compact since Z is compact. Therefore, there exists $z^* \epsilon S$ such that $f_1(z^*) \leq f_1(z)$ for all $z \epsilon S$. Suppose

$f_0(z^*) \neq k$. Then $z^* \epsilon U = \{z \colon f_0(z) < k\}$ and clearly U is open. Since Z is locally convex, it follows that there exists an $\alpha > 0$ such that $(1-\alpha)z^* + \alpha z_1 \epsilon U \subset S$. Since f_1 is convex,

$$f_1((1-\alpha)z^* + \alpha z_1) \leq (1-\alpha)f_1(z^*) + \alpha f_1(z_1) < (1-\alpha)f_1(z^*) + \alpha f_1(z^*) = f_1(z^*)$$

which is a contradiction to the definition of z^*.

We will now prove the first useful convexity result of this article. We consider the family of priors (1.1) and let Σ_1 and Σ_2 be multiples of the identity matrix. This extra condition will not be very restrictive because if it was thought that $\Sigma_1 \leq \Sigma \leq \Sigma_2$ where Σ_1 and Σ_2 are not necessarily multiples of the identity, we could always augment the set of variance–covariance matrices by using the obvious fact that any nonnegative definite matrix is bounded below and above (in the sense we have been talking about) by multiples of the identity matrix. Such an augmentation would not be very conservative unless in the subjective elicitation process we end up with Σ_1 and Σ_2 with very spread out eigenvalues. It seems that such a thing is unlikely because the general tendency would be to keep the subjective elicitation simple and therefore one would probably use diagonal bounds on Σ anyway, adjusting the diagonal elements until it is thought that sufficiently high nonzero correlations have been allowed by using these bounds (e.g., if the bounds $I \leq \Sigma \leq 5I$ are used in 2 dimension, then the correlation is between $\pm\frac{2}{3}$). So eventually it will come down to replacing the diagonal elements by their minimum and maximum respectively. This may be bad if the opinions on some of the regression coefficients are much more precise than those on the other coefficients. Even then, we are only suggesting augmenting the family of priors and therefore a more cautious analysis. The condition that Σ_1 and Σ_2 are multiples of the identity seems crucial for the convexity result to hold.

Theorem 2.2. Let $\utilde{Y}_{n\times 1} \sim N(X\utilde{\theta}, \sigma^2 I)$, where $\utilde{\theta} \in \mathbb{R}^p$ is unknown and $\sigma^2 > 0$ is known. Let $\utilde{\theta}$ have the family of priors Γ as in (1.1) with $\Sigma_1^{-1} = kI$ and $\Sigma_2^{-1} = \ell I$, $k > \ell > 0$. Suppose it is desired to estimate $\utilde{\theta}_{p\times 1}$ under the loss

$$L(\utilde{\theta}, \utilde{a}) = \|\utilde{\theta} - \utilde{a}\|^2. \tag{2.1}$$

Let $r(\Sigma, \delta_\Sigma)$ denote the Bayes risk of $\delta_\Sigma(\utilde{y}) = (X'X + \Sigma^{-1})^{-1}(X'\utilde{y} + \Sigma^{-1}\utilde{\mu})$ with respect to the $N(\utilde{\mu}, \Sigma)$ prior (note δ_Σ is the Bayes rule for this prior when the design matrix X

is used). Then $\sup\limits_{k^{-1}I\leq\Sigma\leq\ell^{-1}I} r(\Sigma,\delta_\Sigma) - \inf\limits_{k^{-1}I\leq\Sigma\leq\ell^{-1}I} r(\Sigma,\delta_\Sigma)$ is a decreasing and convex functional (on the space of nonnegative definite matrices).

The proof of Theorem 2.2 needs the following lemma.

Lemma 2.3. Let R, S, T be symmetric nonnegative definite matrices such that $ST = TS$. Then $tr(RST) \geq 0$.

Proof: See appendix.

Proof of Theorem 2.2: Let $X'X = M$. Assume without loss $\sigma^2 = \ell = 1$. Since $r(\Sigma,\delta_\Sigma) = tr(M+\Sigma^{-1})^{-1}$, it follows that the range of the Bayes risks equals $tr(M+I)^{-1} - tr(M+kI)^{-1}$. In order to prove that the functional

$$\Phi(M) = tr(M+I)^{-1} - tr(M+kI)^{-1} \tag{2.2}$$

is convex in M, familiar arguments imply that it is enough to prove that for a fixed *nnd* matrix M and a fixed symmetric matrix P such that $M+\alpha P$ is *nnd*, the real valued function

$$g(\alpha) = tr(M+\alpha P+I)^{-1} - tr(M+\alpha P+kI)^{-1} \tag{2.3}$$

is convex in the real variable α for $0 \leq \alpha \leq 1$. Define $Q = M+\alpha P+I$. Direct computation yields

$$g''(\alpha) = 2\left[trQ^{-1}PQ^{-1}PQ^{-1} - tr(Q+(k-1)I)^{-1}P(Q+(k-1)I)^{-1}P(Q+(k-1)I)^{-1}\right] \tag{2.4}$$

$$\begin{aligned} \text{Let } Q^{-1} &= A \\ (Q+(k-1)I)^{-1} &= B, \\ PQ^{-1}P &= C, \\ \text{and } P(Q+(k-1)I)^{-1}P &= D. \end{aligned}$$

Then, using (2.4), it suffices to show that $trACA \geq trBDB$.

$$\begin{aligned} \text{Now, } trACA - trBDB &= tr(A-B)C(A+B) + tr(BCA-ACB) + trB(C-D)B \\ &= trC(A+B)(A-B) + tr(BCA-ACB) + tr(C-D)BB. \end{aligned} \tag{2.5}$$

Now note that A, B, C, D are symmetric *nnd*, $A \geq B$, $C \geq D$, and $AB = BA$. Hence, $(A+B)$ and $(A-B)$ commute and $tr(BCA) = tr(ACB)$, implying by virtue of Lemma 2.3 that $g''(\alpha) \geq 0$.

To show that $\Phi(M)$ is decreasing it is enough to show that $g'(\alpha) \leq 0$ whenever P is *nnd*. However

$$g'(\alpha) = -trAPA + trBPB$$

and from Lemma 2.3 (or the previous argument with $C = D = P$) we have $g'(\alpha) \leq 0$. This proves the theorem.

Suppose now we seek a design that minimizes the range of the Bayes risks subject to the restriction of being $\in$–Bayes with respect to a fixed $N(\underset{\sim}{\mu}, \Sigma_0)$ prior; this later restriction will be of the form $tr(X'X + \Sigma_0^{-1})^{-1} \leq K$. Since $\Phi_1(M)$, the range of the Bayes risks, is convex, and $tr(X'X + \Sigma^{-1})^{-1}$ is continuous (and convex), Theorem 2.1 will apply provided the set of possible information matrices is compact and typically the optimal design will be an $M = M^*$ for which $tr(M^* + \Sigma_0^{-1})^{-1} = K$. There usually also will be the very convenient reduction in that the functional $\Phi_1(M)$ is decreasing in M (in the familiar sense) so that we will often need to consider only the information matrices "on the boundary." The problem will then simplify to finding an information matrix M^* on the boundary with the property $tr(M^* + \Sigma_0^{-1})^{-1} = K$. This will, in simple cases, reduce to a problem in just a few real variables. We will explicitly demonstrate the optimal design for linear regression in the case when the independent variable x takes values in $[-1, 1]$. We will also give a numerical example involving the estimation of p treatment means.

Example 1. Consider the simple linear regression model $EY = \theta_0 + \theta_1 x$, where $-1 \leq x \leq 1$. Suppose $\underset{\sim}{\theta} = (\theta_0, \theta_1)'$ has a $N(\underset{\sim}{\mu}, \sigma^2\Sigma)$ distribution and suppose $\ell I \leq \Sigma^{-1} \leq kI$ for some ℓ and $k (0 < \ell < k)$. Let $\Sigma_0^{-1} = R = \begin{pmatrix} r_1 & r \\ r & r_2 \end{pmatrix}$ be any fixed matrix (usually in the above range). We want to find the design minimizing $\Phi_1(M) = tr(M + \ell I)^{-1} - tr(M + kI)^{-1}$ subject to $\Phi_0(M) = tr(M + \Sigma_0^{-1})^{-1} \leq (1+\epsilon) tr(M_0 + \Sigma_0^{-1})^{-1}$ where M_0 is the Bayes design under Σ_0, i.e. M_0 minimizes $\Phi_0(M)$. Here n observations will be taken on the response y and $\epsilon > 0$ is a fixed number.

Since Φ_0 and Φ_1 are both decreasing we can, by familiar arguments, restrict attention to two point designs that sample only at $x = \pm 1$. The matrices M under consideration are thus of the form

$$M = n \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix} \tag{2.6}$$

where $|c| \leq 1$. It is easy to check that the design M_0 (i.e. the Bayes design under Σ_0) is given by $c = -r/n$ provided $|r| \leq n$. If $\ell I \leq R \leq kI$, this will be the case if $n \geq (k - \ell)/2$. Also, the Bayes risk under Σ_0 equals

$$\Phi_0(M_0) = tr(M_0 + R)^{-1} = \frac{2n + r_1 + r_2}{(n + r_1)(n + r_2)}. \tag{2.7}$$

Then $\Phi_0(M) = (1 + \epsilon)\Phi(M_0)$ iff

$$(nc + r)^2 = \frac{\epsilon}{1 + \epsilon}(n + r_1)(n + r_2). \tag{2.8}$$

Obviously (2.8) has two solutions in c. Since $\Phi_1(M)$ is convex in c and symmetric about zero its minimum is at $c_1 = 0$. Consequently our required minimum c^* is the root of (2.8) that is closer to zero, provided $c = 0$ does not already satisfy the constraint. Therefore

$$\begin{aligned} c^* &= \frac{1}{n}\left(-r + \sqrt{\frac{\epsilon}{1+\epsilon}(n + r_1)(n + r_2)}\right) \text{ if } r \geq 0 \\ &= \frac{1}{n}\left(-r - \sqrt{\frac{\epsilon}{1+\epsilon}(n + r_1)(n + r_2)}\right) \text{ if } r < 0 \end{aligned} \tag{2.9}$$

One can check that $c = 0$ already satisfies the constraint if

$$\epsilon \geq \epsilon_0 = \frac{r^2}{(n + r_1)(n + r_2) - r^2} \tag{2.10}$$

Thus if $\epsilon \geq \epsilon_0$ the solution is $c^* = 0$; otherwise c^* is given by (2.9).

Before discussing example 2 we give a brief discussion of some Lagrangian theory which is intimately related to the Kiefer–Wolfowitz equivalence theory. Since $\Phi_0(M) = tr(M + \Sigma^{-1})^{-1}$ is convex in M, the information matrix M_0 minimizes $\Phi_0(M)$ iff $g_0(\alpha) = \Phi_0(\alpha M_0 + (1 - \alpha)M)$ satisfies $g_0'(0) \geq 0$ for all M. Since

$$g_0'(0) = -tr(M_0 + \Sigma^{-1})^{-1}(M_0 - M)(M_0 + \Sigma^{-1})^{-1}$$

and all M are of the form $M = n\Sigma p_i x_i x_i'$, we find that M_0 minimizes $\Phi_0(M)$ iff

$$\underset{\sim}{x}'(M_0+\Sigma^{-1})^{-2}\underset{\sim}{x} \leq C_0 \text{ for all } \underset{\sim}{x} \in \mathcal{X}, \tag{2.11}$$

where $nC_0 = tr(M_0+\Sigma^{-1})^{-2}M_0$. Moreover equality must hold for $\underset{\sim}{x}_i$ used in M_0. In minimizing $\Phi_1(M) = tr(M+\ell I)^{-1} - tr(M+kI)^{-1}$ subject to $\Phi_0(M) \leq (1+\epsilon)\Phi_0(M_0)$ it is fairly easy to show that M^* is the minimum iff there is a $u \geq 0$ such that $g_1'(0)+ug_0'(0) \geq 0$ for all M (where $g_i(\alpha) = \Phi_i((1-\alpha)M^* + \alpha M))$ and $u(\Phi_0(M^*) - (1+\epsilon)\Phi_0(M_0)) = 0$. If $u = 0$ the global minimum of Φ_1 results and the constraint on Φ_0 is inactive. For the functional Φ_1 at hand, we find that M^* is the constrained minimum if for some $u > 0$

$$\underset{\sim}{x}'(M^*+\ell I)^{-1}\underset{\sim}{x} - \underset{\sim}{x}'(M^*+kI)^{-1}\underset{\sim}{x} + u\underset{\sim}{x}'(M^*+\Sigma^{-1})^{-1}\underset{\sim}{x} \leq C^* \tag{2.12}$$

where

$$nC^* = tr(M^*+\ell I)^{-2}M^* - tr(M^*+kI)^{-2}M^* + u\ tr(M^*+\Sigma^{-1})^{-2}M^*, \tag{2.13}$$

and

$$\Phi_0(M^*) = (1+\epsilon)\Phi_0(M_0). \tag{2.14}$$

Again equality must occur in (2.12) for any $\underset{\sim}{x}_i^*$ in $M^* = \Sigma p_i^*(x_i^*)(x_i^*)'$ for which $p_i^* > 0$.

The above results are called "equivalence" theorems since, for example, the minimization of $\Phi_0(M)$ is equivalent to (2.11). An elaborate literature on such theorems is available. See, for example, Pukelsheim and Titterington (1983), and Gaffke (1985). A short discussion of the equivalence theory for constrained problems is given in Lee (1988). For historical record we should remark here that a form of the original Kiefer–Wolfowitz theorem is given in Schoenberg (1959) in the original context of D–optimality, i.e. finding the design which maximizes the determinant of $X'X$.

Example 2. Consider a completely randomized design with p treatments and suppose the treatment means $\mu_1, \mu_2, \ldots, \mu_p$ have a prior as in (1.1) and again suppose $\ell I \leq \Sigma^{-1} \leq kI$. The information matrix M is now a diagonal matrix with diagonal elements n_i = number of measurements on μ_i. For the case $p = 2$ a complete solution is easily given as in Example

1 for general $\Sigma_0^{-1} = \begin{pmatrix} r_1 & r \\ r & r_2 \end{pmatrix}$. We omit these details. For arbitrary p, assume that $\Sigma_0^{-1} = R_0$ is diagonal with diagonal elements r_i and assume without loss of generality that $0 < r_1 \leq r_2 \leq \ldots \leq r_p$. Measurements on μ_i correspond to $\underset{\sim}{x}_i = \underset{\sim}{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ with the "one" in the i^{th} component. From (2.11) it follows that M_0 minimizes $\Phi_0(M)$ iff $(n_i^0 + r_i)^{-2} = C_0$ whenever $n_i^0 > 0$ and $r_i^{-2} \leq C_0$ if $n_i^0 = 0$. It is easy to see that if $n \geq pr_p - \Sigma r_i$, then $n_i^0 + r_i^0 \equiv \lambda_0 = \frac{n + \Sigma r_i}{p}$ for all i. Note $n_1^0 \geq \ldots \geq n_p^0$. Intuitively, one makes the posterior precisions $n_i + r_i$ as equal as possible (starting with the smallest r_i).

Since the functional Φ_1 is convex and invariant under permutations of the treatments it follows that the minimum of Φ_1 occurs for $n_i = n/p$. The minimum of Φ_1 subject to $\Phi_0(M) \leq (1 + \epsilon)\Phi_0(M_0)$ amounts to moving the n_i from n_i^0 in the "direction" of n/p. Equation (2.12) shows that $n_i^*, i = 1, 2, \ldots, p$ is the required solution if (2.14) holds and for some $u > 0$

$$(n_i^* + \ell)^{-2} - (n_i^* + k)^{-2} + u(n_i^* + r_i)^{-2} = C^* \qquad (2.15)$$

where C^* is given by (2.13). The condition on C^* in (2.13) will force $\Sigma n_i^* = n$. In solving these equations we actually solve (2.15), (2.14) and $\Sigma n_i^* = n$ for u, C^* and $n_1^*, \ldots, n_p^*$.

Two examples are illustrated in Figures 2.1 and 2.2. Fig. 2.1 corresponds to $\ell = 1$ and $k = 5$ and $n = 15$ while Fig. 2.2 has $\ell = 1$ and $k = 9$ and $n = 25$. In Fig. 2.1 and Fig. 2.2, the other parameters are given by the following table; quantities in parentheses apply to Fig. 2.2.

	$p=2$	$p=3$	$p=5$
r_1	1(1)	1(1)	1(1)
r_2	5(9)	3(5)	2(3)
r_3		5(9)	3(5)
r_4			4(7)
r_5			5(9)

The value plotted in the two figures is

$$\eta(\epsilon) = (\Phi_1^0 - \Phi_1^*)/\Phi_1^0$$

where Φ_1^0 is the value of Φ_1 at the minimum for Φ_0, and Φ_1^* is the value at the constrained minimum. Thus $100\eta(\epsilon)$ is the percent gain in robustness for a sacrifice of 100 ϵ % in subjective Bayes risk. We remark, and it is not very hard to show, that for ϵ near zero, the value of $\eta(\epsilon)$ is approximately

$$\eta(\epsilon) \approx \frac{(n+\Sigma r_i)s_a}{\Phi_1^0}\sqrt{\epsilon}$$

where $s_a^2 = p^{-1}\sum_{i=1}^{p}(a_i - \bar{a})^2$, $\bar{a} = \Sigma a_i/p$, and $a_i = -(n_i^0 + \ell)^{-2} + (n_i^0 + k)^{-2}$. Thus, the percentage gain is considerable for small ϵ. As an example for $n = 25$, (see fig. 2.2) $p = 2, r_1 = \ell = 1, r_2 = k = 9, \epsilon = .02$ corresponds to $\eta(\epsilon) = .14$, which represents a 14% gain in robustness for a 2% sacrifice in risk. At this point n_1 and n_2 have moved from $n_1^0 = 16.5$ and $n_2^0 = 8.5$ (where $n_1^0 + r_1 = n_2^0 + r_2 = 17.5$) to $n_1^* = 14$ and $n_2^* = 11$. For fixed n, the constant multiplying $\sqrt{\epsilon}$ appears (as in the figures) to be increasing in p. This provides further confirmation, that there is generally more gain in robustness for fixed ϵ, for larger values of p, i.e. more parameters in the model. For $n = 15$ the constants are approximately 1.3, 1.7 and 2.6 for $p = 2$, 3 and 5 respectively.

Before continuing with our general discussion we consider some counterexamples which illustrate that the matrix functional

$$\rho(M) = (M + R_2)^{-1} - (M + R_1)^{-1} \tag{2.16}$$

is neither decreasing nor convex in M. Here $R_2 = \Sigma_2^{-1}, R_1 = \Sigma_1^{-1}, R_1 \geq R_2 > 0$ so that (2.16) is *nnd*. In general $(M + R_i)^{-1}$ is decreasing and convex in M. Let $A_i = (M + R_i)^{-1}$ for $i = 1, 2$. By considering the differential of (2.6) in α for $M + \alpha P$ we see that the difference (2.16) is decreasing in M iff

$$-A_2PA_2 + A_1PA_1 \leq 0 \tag{2.17}$$

for all $P \geq 0$. Moreover the difference will be convex iff

$$A_2PA_2PA_2 - A_1PA_1PA_1 \geq 0 \tag{2.18}$$

for arbitrary symmetric P.

The two inequalities (2.17) and (2.18) can be violated as the following examples show. Let $A_2 = \begin{pmatrix} 20 & 0 \\ 0 & 2 \end{pmatrix}$, $A_1 = \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix}$ and $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Then $P > 0$ and $0 < A_1 < A_2$.

However $A_2PA_2 = \begin{pmatrix} 400 & 0 \\ 0 & 4 \end{pmatrix}$ and $A_1PA_1 = \begin{pmatrix} 100 & 33 \\ 33 & 10 \end{pmatrix}$ so that (2.17) is violated. Similarly (2.18) is violated if $A_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $P = \begin{pmatrix} 0 & 1 \\ 1 & d \end{pmatrix}$ and $d > 2$.

Since $M = n\Sigma_1 p_i X_i X_i'$ is essentially arbitrary it is easy to see that one can construct linear models for which $\rho(M)$ is neither increasing nor convex. These examples indicate that each functional of $\rho(M)$ must be handled separately.

Sometimes, it may be more desirable to keep the estimates themselves as close as possible rather than keeping the range of the Bayes risks small. As mentioned in the introduction, the collection of Bayes estimates as the prior varies in a class Γ usually form a nice convex set. We can then seek a design that keeps the diameter (in some metric) of this set small; for instance, we could seek a design that keeps the Euclidean diameter (i.e., the diameter under L_2 norm) of this set small. The following result is taken from DasGupta and Studden (1988a). See also Leamer (1978,1982) and Polasek (1984).

Theorem 2.4. Under the setup of Theorem 2.2, the Euclidean diameter of the set S of all Bayes estimates is equal to

$$D = \sqrt{\underset{\sim}{v}'(\Lambda_2 - \Lambda_1)\underset{\sim}{v} \cdot \lambda_{\max}} \ , \tag{2.19}$$

where $\underset{\sim}{v} = X'(\underset{\sim}{y} - X\underset{\sim}{\mu}), \Lambda_1 = (X'X + kI)^{-1}, \Lambda_2 = (X'X + \ell I)^{-1}$, and $\lambda_{\max}$ is the maximum eigenvalue of $\Lambda_2 - \Lambda_1$.

For an expression of D when Σ_1, Σ_2 are arbitrary *nnd* matrices, see DasGupta and Studden (1988a). The problems in working with the Euclidean diameter D are that D depends on $\underset{\sim}{y}$ and consequently an expected value has to be taken to address a design problem. It turns out that $E(D^2)$ is usually easier to calculate and handle than $E(D)$. If we take an expected value of D^2 under the marginal distribution of $\underset{\sim}{y}$ induced by a $N(\underset{\sim}{\mu}, \Sigma)$ prior, then only in very special cases it can be proved to be convex. However, in Theorem 2.1 we noted that for our restricted optimal design problems we can sometimes bypass the question of convexity by letting the nonconvex functional as f_0 and another appropriate functional that we know is convex as f_1. For example, the Bayes risk under a $N(\underset{\sim}{\mu}, \Sigma)$ prior equals the functional $tr(M + \Sigma^{-1})^{-1}$ and this is known to be decreasing and convex. In such a case, we can reformulate our design problem as minimizing $tr(M+\Sigma^{-1})^{-1}$ subject to keeping $E(D^2)$ small. Of course, we will still need to establish that $E(D^2)$ is decreasing in M; this is crucial in order to do all the minimizations "on the boundary" of the space of information matrices. As we shall later see, $E(D^2)$ is often decreasing in the important special cases of polynomial regression. First, we give below a result that says that $E(D^2)$ is convex in some cases.

Theorem 2.5. Suppose $\ell = 0$ in Theorem 2.4 (i.e., Σ varies in the range $\frac{1}{k}I \leq \Sigma$). Then the expected value of D^2 under the marginal distribution of $\underset{\sim}{y}$ induced by the $N(\underset{\sim}{\mu}, \frac{1}{k}I)$ prior is decreasing and convex.

Proof: Direct computation gives that the expected value of D^2 is proportional to

$$\begin{aligned}\phi(M) &= \lambda_{\max}\{M^{-1} - (M+kI)^{-1}\}\\ &= \frac{k}{\lambda_s(k+\lambda_s)},\end{aligned}$$

where λ_s is the smallest eigenvalue of M. It's self–evident that ϕ is decreasing. To show that it is convex, as usual it will suffice to show that

$$g(\alpha) = \frac{1}{h(\alpha)[k+h(\alpha)]}$$

is convex in the real variable $\alpha (0 \leq \alpha \leq 1)$ where

$$\begin{aligned}h(\alpha) &= \lambda_s(M+\alpha P)\\ &= \text{ the smallest eigenvalue of } M+\alpha P,\end{aligned}$$

where P is a symmetric matrix such that $M+\alpha P$ is p.d. Since $g(\alpha) > 0$, to prove that g is convex it will suffice to prove that g is log convex (i.e., $\log g$ is convex). Now, $\log g = -[\log h + \log(k+h)]$, and k, $h > 0$. Consequently, it will be enough to prove that $h(\alpha)$ is concave ($h > 0, h$ concave $\Rightarrow \log h$ concave). But now recall that

$$h(\alpha) = \lambda_s(M+\alpha P),$$

and h is concave in α if and only if $\lambda_s(M)$ is concave in M, which is known to be true; this proves the theorem.

Here is an application of this result.

Example 3. In the simple linear regression model, let $\underset{\sim}{\theta} \sim N(\underset{\sim}{\mu}, \Sigma)$, $\Sigma \geq \frac{1}{k}I$. Then the design that minimizes the expected squared diameter of the set of Bayes estimates subject to the restriction $tr(M + \Sigma_0^{-1})^{-1} \leq (1+\varepsilon)tr(M_0 + \Sigma_0^{-1})^{-1}$ where Σ_0 is any fixed matrix and M_0 the corresponding Bayes design is again given by the root of (2.8) that is closer to zero. This is because the global optimum of $E(D)^2$ is attained at $c = 0$. Observe the similarity in the optimal design in examples 1 and 3.

Example 4. Again, consider the completely randomized design for estimating p means $\mu_1, \ldots, \mu_p$ discussed in example 2. The case $p = 2$ can again be handled quite generally. Here we would like to minimize the functional $\Phi_2(M) = \lambda_{\max}(M^{-1}(M + kI)^{-1})$, again subject to $\Phi_0(M) \leq (1+\epsilon)\Phi_0(M_0)$ where $\Sigma_0^{-1} = R$ is diagonal as in example 2. We assume that $r_1 \leq r_2 \leq \ldots \leq r_p$ so that $n_1^0 \geq \ldots \geq n_p^0$. Thus $\Phi_2(M_0) = \frac{1}{n_p^0(n_p^0+k)}$. For illustrative purposes we assume $r_{p-1} < r_p$ and n is large enough so $0 < n_p^0 < n_{p-1}^0$. For ϵ sufficiently small we can apply (2.12) to show (or show directly) that the constrained solution n_i^* satisfies $n_p^* + r_p = \lambda_0''$ and $n_i^* + r_i = \lambda_0'$, $i = 2, \ldots, p$ where λ_0' and λ_0'' are determined by $\Sigma n_i^* = n$ and the constraint $\Phi_0(M^*) = (1+\epsilon)\Phi_0(M_0)$. The general solution is to set

$$n_i^* + r_i = \lambda_0'' \text{ for } i \leq i_0 \text{ and } \lambda_0' \text{ for } i > i_0$$

for some i_0 depending on ϵ. We omit the details.

In the context of polynomial regression, special interest lies in estimating the coefficient of the highest order term; this is a parametric function of the form $\underset{\sim}{c}'\underset{\sim}{\theta}$ where $\underset{\sim}{c} = (0\ 0 \ldots 0\ 1)'$. The Bayes risk under an arbitrary $N(\underset{\sim}{\mu}, \Sigma)$ prior for this problem (assuming squared error loss) is $\underset{\sim}{c}'(M + \Sigma^{-1})^{-1}\underset{\sim}{c}$, which is convex (actually, for any $\underset{\sim}{c}$). We prove below that the range of the Bayes risks is decreasing and consequently, Theorem 2.1 again applies and it is possible to identify the design that minimizes the Bayes risk with respect to an arbitrary Σ subject to the range of the Bayes risks being sufficiently small. A nice feature of our next result is that for arbitrary *nnd* matrices Σ_1 and Σ_2 in (1.1), the range of the Bayes risks can be proved to be decreasing in M.

Theorem 2.6. Consider the polynomial regression model $y_i = \theta_0 + \sum_{j=1}^{p} \theta_j x_i^j + \varepsilon_i$, where ε_i are *iid* $N(0, \sigma^2)$. Consider the problem of estimating θ_p under squared error loss. Suppose $\underset{\sim}{\theta}$ has a prior as in (1.1). Then the Bayes risk under any specific Σ is decreasing and convex and the range of the Bayes risks is decreasing.

Proof: The Bayes risk under a specific Σ is $\underset{\sim}{c}'(M + \Sigma^{-1})^{-1}\underset{\sim}{c}$ where $\underset{\sim}{c} = (0\ 0 \ldots 0\ 1)'$ and this is known to be decreasing and convex (Chaloner (1984)). On the other hand, the range of the Bayes risks equals

$$\begin{aligned} \Phi(M) =& \underset{\sim}{c}'(M + \Sigma_2^{-1})^{-1}\underset{\sim}{c} - \underset{\sim}{c}(M + \Sigma_1^{-1})^{-1}\underset{\sim}{c} \\ &= \text{The diagonal element in the bottom corner of} \\ & (M + \Sigma_2^{-1})^{-1} - (M + \Sigma_1^{-1})^{-1}. \end{aligned}$$

Let now

$$\Sigma_1^{-1} = \begin{pmatrix} B_1 & \underset{\sim}{\mathrm{v}}_1 \\ \underset{\sim}{\mathrm{v}}_1' & \alpha_1 \end{pmatrix},$$
$$\text{and } \Sigma_2^{-1} = \begin{pmatrix} B_2 & \underset{\sim}{\mathrm{v}}_2 \\ \underset{\sim}{\mathrm{v}}_2' & \alpha_2 \end{pmatrix}.$$

Since we have a polynomial regression model, it is well known that in order to prove that $\Phi(M)$ is decreasing in M, we have to only show that if M is partitioned as

$$M = \begin{pmatrix} A & \underset{\sim}{\mathrm{u}} \\ \underset{\sim}{\mathrm{u}}' & x \end{pmatrix},$$

then for fixed $\underset{\sim}{u}$ and A, $\Phi(M)$ is decreasing in the real variable x for $x \geq 0$. Standard matrix identities give that

$$\begin{aligned} \Phi(M) &= \frac{1}{x + \alpha_2 - (\underset{\sim}{u} + \underset{\sim}{v}_2)'(A + B_2)^{-1}(\underset{\sim}{u} + \underset{\sim}{v}_2)} \\ &\quad - \frac{1}{x + \alpha_1 - (\underset{\sim}{u} + \underset{\sim}{v}_1)'(A + B_1)^{-1}(\underset{\sim}{u} + \underset{\sim}{v}_1)}. \end{aligned} \tag{2.19}$$

Now recall that $\Sigma_2^{-1} < \Sigma_1^{-1}$

$$\begin{aligned} &\Rightarrow C_2 = \left(\begin{pmatrix} A & \underset{\sim}{\mathrm{u}} \\ \underset{\sim}{\mathrm{u}}' & 0 \end{pmatrix} + \Sigma_2^{-1} \right)^{-1} > C_1 = \left(\begin{pmatrix} A & \underset{\sim}{\mathrm{u}} \\ \underset{\sim}{\mathrm{u}}' & 0 \end{pmatrix} + \Sigma_1^{-1} \right)^{-1} \\ &\Rightarrow \text{The diagonal element in the bottom corner of } C_2 - C_1 \text{ is positive} \\ &\Rightarrow \alpha_2 - (\underset{\sim}{u} + \underset{\sim}{v}_2)'(A + B_2)^{-1}(\underset{\sim}{u} + \underset{\sim}{v}_2) < \alpha_1 - (\underset{\sim}{u} + \underset{\sim}{v}_1)'(A + B_1)^{-1}(\underset{\sim}{u} + \underset{\sim}{v}_1) \end{aligned} \tag{2.20}$$

(note C_1, C_2 are both p.d.);

(2.19) and (2.20) now imply that $\Phi(M)$ is decreasing in x. This proves the Theorem. An example follows.

Example 5. Consider again the simple linear model $E(y) = \theta_0 + \theta_1 x$ and suppose $-1 \leq x \leq 1$. In this case, letting $\Sigma_2^{-1} = \ell I$ and $\Sigma_1^{-1} = kI$, (2.19) reduces to

$$\Phi(M) = \frac{k-\ell}{(n+\ell)^2(n+k)^2} \cdot \frac{1+\alpha\beta c_1^2}{(1-\alpha^2 c_1^2)(1-\beta^2 c_1^2)}, \tag{2.21}$$

where $\alpha = \frac{n}{n+\ell}, \beta = \frac{n}{n+k}$, and

$$\begin{pmatrix} n & nc_1 \\ nc_1 & n \end{pmatrix} = M; \text{ notice } -1 \leq c_1 \leq 1.$$

It is straightforward to check that the function $\frac{1+\alpha\beta c_1^2}{(1-\alpha^2 c_1^2)(1-\beta^2 c_1^2)}$ is log convex for $1 \geq c_1 \geq 0$ and hence convex for $-1 \leq c_1 \leq 1$ (that $1 > \alpha > \beta > 0$ is needed to check this). The Bayes risk of an arbitrary $M = \begin{pmatrix} n & nc_1 \\ nc_1 & n \end{pmatrix}$ with respect to an arbitrary Σ, where $\Sigma^{-1} = \begin{pmatrix} r_1 & r \\ r & r_2 \end{pmatrix}$, is given by

$$r(\Sigma) = \frac{n+r_1}{(n+r_1)(n+r_2)-(nc_1+r)^2}, \tag{2.22}$$

and the global minimum value of $r(\Sigma)$ over M equals $\frac{1}{n+r_2}$. Combining now the convexity (and decreasingness) of (2.19) and (2.22), the design minimizing (2.19) subject to $r(\Sigma) \leq \frac{1+\epsilon}{n+r_2}$, is the root of the equation

$$\frac{n+r_1}{(n+r_1)(n+r_2)-(nc_1+r)^2} = \frac{1+\varepsilon}{n+r_2} \tag{2.23}$$

closer to $c_1 = 0$ (which is the global minima of (2.21)).

Remark 3. For higher order models, it is hard to show that the range of the Bayes risks $\underset{\sim}{c}'(M+\Sigma_2^{-1})^{-1}\underset{\sim}{c} - \underset{\sim}{c}'(M+\Sigma_1^{-1})^{-1}\underset{\sim}{c}$ is convex. In these cases, Theorems 2.1 and 2.6 still enable us to restrict attention to matrices M on the appropriate boundaries which make $\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{c}$ exactly equal to the imposed upper bound. A search can then be made to locate the design minimizing the range among these restricted designs. For example, for quadratic regression our Theorems 2.1 and 2.5 enable us to reduce the optimal design problem to a two variable minimization problem by considering designs of the form

$$M = n\begin{pmatrix} 1 & c_1 & c_2 \\ & c_2 & c_3 \\ & & c_4 \end{pmatrix},$$

$$\begin{aligned} \text{where } c_1 &= (p_2 - p_1) + p_3 a, \\ c_2 &= p_2 + p_1 + p_3 a^2, \\ c_3 &= (p_2 - p_1) + p_3 a^3 \\ \text{and } c_4 &= p_2 + p_1 + p_3 a^4, \end{aligned}$$

where $-1 < a < 1$, and $p_1 + p_2 + p_3 = 1$ (ordinarily, this is a three dimensional minimization; but Theorem 2.1 reduces the dimension by one more).

Remark 4. Again, sometimes it may be more desirable to control the diameter (length) of the interval of Bayes estimates of $\underset{\sim}{c}'\underset{\sim}{\theta}$ instead of controlling the range of the Bayes risks. From corollary 2.2(b) in DasGupta and Studden (1988a), it follows that if $\Sigma \geq \frac{1}{k}I$ in (1.1), then the expected squared length (under the marginal distribution of $\underset{\sim}{y}$ induced by the $N(\underset{\sim}{\mu}, \frac{1}{k}I)$ prior) of the interval of Bayes estimates of $\underset{\sim}{c}'\underset{\sim}{\theta}$ is, upto a proportionality constant, $\underset{\sim}{c}'M^{-1}\underset{\sim}{c} - \underset{\sim}{c}'(M + kI)^{-1}\underset{\sim}{c}$; this is decreasing by Theorem 2.6 (and convex for the simple linear regression case) so that the usual argument laid out in the above will again apply.

3. Priors inside a density band. In this section, we consider construction of optimum designs when we have the family of priors (1.2) with L a $N(\underset{\sim}{\mu}, \sigma^2\Sigma)$ density and $U = kL$ where $k > 1$. As k gets larger, the family of priors (1.2) also gets larger. As we will like to concentrate on the experimental design issues in this article, we refrain from giving an extensive discussion of this family of priors and instead refer the reader to DeRobertis (1978), DeRobertis and Hartigan (1981), Berger (1987), and sections 1 and 6 in DasGupta and Studden (1988a). However, we remind the reader that in contrast to the family of priors (1.1), the mean and the variance–covariance structure all change simultaneously as the prior changes in the family (1.2). To give the reader a flavor of how different the prior

means can be, we consider the model $Y \sim N(X\theta, \sigma^2 I)$ when $\sigma^2 = 1$ and L is $N(\underset{\sim}{\mu}, I)$. The prior mean of each θ_i is in the range $\mu_i \pm \gamma(k)$, where $\gamma(k)$ for various values of k is given below:

k	2	3	4	5	6	8	10
$\gamma(k)$	.276	.436	.549	.636	.707	.817	.901

Thus, for example, if θ_1 has mean zero and variance 1 under L, then the prior mean varies between $\pm .549$ for $k = 4$. The nice feature of our results in this section is that the design which is Bayes with respect to L will be seen to have a number of robustness properties as well. The following Theorem is taken from DasGupta and Studden (1988a).

Theorem 3.1. Consider the normal linear model $\underset{\sim}{Y}_{n\times 1} \sim N(X\underset{\sim}{\theta}, \sigma^2 I)$, where $\underset{\sim}{\theta}_{p\times 1}$ is unknown and $\sigma^2 > 0$ is known. Let $\underset{\sim}{\theta}$ have a prior belonging to the family (1.2) with L as a $N(\underset{\sim}{\mu}, \sigma^2\Sigma)$ density and $U = kL,\ k > 1$. Then the Euclidean diameter of the set of Bayes estimates of $\underset{\sim}{\theta}$ for squared error loss is free of $\underset{\sim}{y}$ and equals

$$D_L = 2\gamma \cdot \sqrt{\lambda_{\max}(M + \Sigma^{-1})^{-1}}\ , \tag{3.1}$$

where γ is an absolute constant depending on k.

The very attractive feature of the above theorem is that D_L is independent of $\underset{\sim}{y}$ and therefore unlike in section 2, we do not need to take an expected value of D_L (or its square). The idea here is that if at the design stage we somehow knew what the $\underset{\sim}{y}$ data would be, then a Bayesian design should be geared towards optimum performance for this fixed data. A value of D_L independent of $\underset{\sim}{y}$ enables us to do exactly that.

Recall now that the family of priors in Theorem (3.1) is a metric neighborhood of the prior L. Consequently, L is a natural choice for the specific prior with respect to which one would like to be nearly Bayes. Since L is a $N(\underset{\sim}{\mu}, \sigma^2\Sigma)$ prior, the Bayes risk with respect to L is, upto a proportionality factor, simply $tr(M + \Sigma^1)^{-1}$. Our restricted optimization problem would then be to minimize $tr(M + \Sigma^{-1})^{-1}$ subject to keeping D_L small, or equivalently $\lambda_{\max}(M + \Sigma^{-1})^{-1}$ small (or vice versa). Since both of these functionals are decreasing and convex, we have a relatively neat scenario in this case. In general, the family of priors (1.2) does give such neat results.

Example 6. Consider the simple linear model again and suppose that the information on $\underset{\sim}{\theta}$ is vague; so we will take $L \equiv 1$ and $U \equiv k > 1$ (formally, this amounts to taking "$\Sigma = \infty$" in Theorem 3.1). Thus the Bayes risk under L equals trM^{-1} and D_L^2 (but for a constant) equals $\lambda_{\max}M^{-1}$. Assume $-1 \leq x \leq 1$; then the global minimum of trM^{-1} is $\frac{2}{n}$ and is attained by the design $M = nI$. On the other hand, the global minimum of $\lambda_{\max}M^{-1}$ is $\frac{1}{n}$ and is also attained at $M = nI$. Therefore, $M = nI$ minimizes $\lambda_{\max}M^{-1}$ subject to $trM^{-1} \leq (1+\epsilon) \cdot \frac{2}{n}$ for every $\varepsilon > 0$. Interestingly, thus, the standard A (and E)–optimal design is noninformative Bayes and also is the solution to the robust design problem.

Before we proceed to give the next example, we would like to point out that in fact $tr(M + \Sigma^{-1})^{-1}$ and $\lambda_{\max}(M + \Sigma^{-1})^{-1}$ are both minimized by the same M in the simple linear model for every Σ.

Example 7. Consider a completely randomized design with p treatments considered in example 2. Let L again be $N(\mu, \sigma^2\Sigma)$ where $\Sigma^{-1} = \text{diag}(r_1, \ldots, r_p)$. The problem here is to minimize $\lambda_{\max}(M + \Sigma^{-1})^{-1}$ subject to $\Phi_0(M) = tr(M + \Sigma^{-1})^{-1}$ being near its minimum. In this example the minimum of both functionals is attained for the same set of $n_1^0, n_2^0, \ldots, n_p^0$. These values are such that $n_i^0 + r_i = \lambda_0$ and are described in example 2. Thus the Bayes risk under the prior L for estimating the vector of treatments and the squared diameter of the set of Bayes estimates are minimized simultaneously.

Example 8. For a quadratic regression model and vague information (i.e., $L \equiv 1$ and $U \equiv k$), suppose we want to minimize D_L^2 subject to a small Bayes risk under L. Usual arguments imply that we can restrict attention to designs of the form

$$M = n \begin{pmatrix} 1 & 0 & c \\ 0 & c & 0 \\ c & 0 & c \end{pmatrix}, \tag{3.2}$$

where $|c| \leq 1$ if $|x| \leq 1$.

The global minimum of trM^{-1} (which is the Bayes risk under L) is attained at $c = \frac{1}{2}$ and the minimum value is 8. On the other hand, the global minimum of $\lambda_{\max}M^{-1}$ (which is proportional to D_L^2) is 5 and is attained at $c = .4$. For $\varepsilon \leq \frac{1}{24}$, the value of c minimizing

D_L^2 subject to $trM^{-1} \leq 8(1+\varepsilon)$ is the lower root of $\frac{2}{c(1-c)} = \frac{8}{1+\varepsilon}$, which is $1-\sqrt{\frac{\frac{\varepsilon}{1+\varepsilon}}{2}}$. For $\varepsilon > \frac{1}{24}$, $c = .4$ satisfies $trM^{-1} \leq 8(1+\varepsilon)$ and hence is the optimum value of c.

As in section 1, we now turn our attention to the estimation of $\underset{\sim}{c}'\underset{\sim}{\theta}$. For the family of priors of this section, we can handle any arbitrary vector $\underset{\sim}{c}$; this enables us to work out the optimal designs for estimation of specific regression coefficients or the mean response at fixed levels of the regressor variables and also for the highly important extrapolation problem in polynomial regression. The following result proved in DasGupta and Studden (1988a) is the central reason that we can handle an arbitrary vector $\underset{\sim}{c}$.

Theorem 3.2. Consider the setup of Theorem 3.1. For any vector $\underset{\sim}{c}$, let $\underset{\sim}{\mu}_c$ and $\underset{\sim}{\sigma}_c$ denote the posterior mean and the posterior standard deviation of $\underset{\sim}{c}'\underset{\sim}{\theta}$ under a fixed prior π, where $L \leq \pi \leq kL$. Define

$$S_{\underset{\sim}{c}} = \{(\underset{\sim}{\mu}_c, \underset{\sim}{\sigma}_c) : L \leq \pi \leq kL\}. \tag{3.3}$$

Then there exists a fixed set S_h, independent of M and $\underset{\sim}{y}$, such that

$$S_{\underset{\sim}{c}} = \sqrt{\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{c}}\; S_h + (\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{v}, 0),$$

where $\underset{\sim}{v} = X'(\underset{\sim}{y} - X\underset{\sim}{\mu})$ and for a real number λ and a fixed vector $\underset{\sim}{\gamma}$, $\lambda A + \underset{\sim}{\gamma}$ denotes the set of all points $\lambda\underset{\sim}{z} + \underset{\sim}{\gamma}$ for $\underset{\sim}{z} \varepsilon A$.

Remark 5. It is actually proved in DasGupta and Studden (1988a) that the fixed set S_h is simply the set

$$S_h = \left\{\left(E(Z), \sqrt{E(Z^2) - E^2(Z)}\right) : Z \sim f,\ \phi \leq f \leq k\phi\right\}, \tag{3.4}$$

where ϕ denotes the standard normal density. In (3.4), $E(Z^k)$ is to be interpreted as $\frac{\int z^k f(z)dz}{\int f(z)dz}$.

Remark 6. Theorem 3.2 implies that the range of the Bayes estimates of $\underset{\sim}{c}'\underset{\sim}{\theta}$ as well as the range of the posterior risks for estimating $\underset{\sim}{c}'\underset{\sim}{\theta}$ can be simultaneously minimized by simply minimizing $\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{c}$ over M. Thus the following theorem is immediate.

Theorem 3.3. Under the setup of Theorem 3.2, the Bayes design with respect to the prior L, i.e., the design minimizing $\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{c}$, also minimizes the range of the Bayes estimates as well as the range of the posterior risks, both of which are independent of the data on the response variable.

Remark 7. An immediate consequence of Theorem 3.3 is that the design minimizing the range of the Bayes estimates (or the range of the posterior risks) subject to an upper bound on the Bayes risk with respect to L is simply the Bayes design under L. Also, this happens for every vector $\underset{\sim}{c}$ and every problem that can be formulated as a normal linear model problem. Several examples follow.

Example 9. The Bayes design minimizing $\boldsymbol{c}'(M+\Sigma^{-1})^{-1}\boldsymbol{c}$ is a rather specialized design specifically built to estimate a specified $\boldsymbol{c}'\boldsymbol{\theta}$. The limiting case $L=1$, $U=kL$ results in the classical designs obtained from minimizing $\boldsymbol{c}'M^{-1}\boldsymbol{c}$. This case has the elegant geometric "Elfving Theorem" associated with it. The result says that if the optimal design has the matrix $M = n\Sigma p_i \underset{\sim}{x}_i \underset{\sim}{x}_i'$ then

$$\Sigma p_i \epsilon_i \underset{\sim}{x}_i = \beta \underset{\sim}{c} \tag{3.5}$$

where $\epsilon_i = \pm 1$ and $\beta\underset{\sim}{c}$ is on the boundary of the convex hull of the set of points $\{\pm\underset{\sim}{x};\underset{\sim}{x}\epsilon\mathcal{X}\}$. The polynomial case where $\underset{\sim}{x} = f(x) = (1,x,\ldots,x^p)'$ for $x\epsilon[-1,1]$ is of particular interest. It is known that if $c_1 = (1,a,\ldots,a^p)'$ for $a>1$ or $c=(0,0,\ldots,0,1)'$ then the optimal design is supported on the "Chebyshev points" $x_\nu = \cos\frac{\nu\pi}{p}, \nu = 0,1,\ldots,p$. These are the zeros of $(1-x^2)T_p'(x)$ where $T_m(x) = \cos m\boldsymbol{\theta}, x = \cos\boldsymbol{\theta}$, is the m^{th} Chebyshev polynomial of the first kind. The weights are found by solving (3.5) with $\epsilon_i = (-1)^i$. Further details in this case can be found in Hoel and Levine (1964), Studden (1968) and Kiefer and Wolfowitz (1965).

In the Bayes context, Chaloner (1984) has observed that for large n the Bayes designs for arbitrary Σ, in the highest coefficient and extrapolation cases, are still supported on the same Chebyshev points. The weights in these cases are found by making

$$\Sigma n_i(-1)^i f(x_i) + R\underset{\sim}{d} = \lambda\underset{\sim}{c} \tag{3.6}$$

for some constant λ. Here $\underset{\sim}{d}$ is the coefficient vector of the polynomial $T_p(x)$ and $R = \Sigma^{-1}$. Chaloner's observation actually holds in more general cases. It is not very hard to show that if $c'\theta$, in any classical situation, has a design supported by a full set $T = T(\underset{\sim}{c})$ of points (i.e. equal in number to the dimension of $\underset{\sim}{\theta}$), then for large n the Bayes design for arbitrary Σ is supported on the same set. Theorem 3.3 therefore implies that the design minimizing the range of the Bayes estimates subject to an upper bound on the Bayes risk under L is supported on the set T. For large n, it is to be expected that the Bayes design will "converge weakly" to the classical design. The surprising observation is that it is supported precisely on the same set for large n.

Example 10. In this example consider weighted polynomial regression on $[-1,1]$. Here we take $\underset{\sim}{x} = f(x) = \sqrt{w(x)}(1, x, \ldots, x^p)$ for $x \epsilon [-1,1]$. The scenario is equivalent to ordinary polynomial regression with variance $\sigma^2\lambda(x)$ where $\lambda(x) = 1/w(x)$. For example, if $w(x) = 1 - x$, then observations have larger variance for x near $x = +1$. Numerous examples for the extrapolation and highest coefficient problems are given in Lau (1983). We quote here one specific result. Thus if $w(x) = 1 - x$ then the optimal classical design for extrapolation or highest coefficient is supported on the zeros of $(1+x)P_p^{(\frac{1}{2},-\frac{1}{2})}(x) = 0$. Here $P_m^{(\alpha,\beta)}(x)$ is the m^{th} Jacobi polynomial orthogonal with weight function $(1-x)^\alpha(1+x)^\beta$. A brief description for the linear case $p = 1$ is given below. Our regression functions are $\sqrt{1-x}(1, x)$. The support of the design is on $x = -1$ and $x = \frac{1}{2}$. The optimal weights can be found from (3.5) and are $\frac{1}{3}$ and $\frac{2}{3}$ respectively. As remarked previously for n large the Bayes design is support on $x = -1$ and $\frac{1}{2}$. If $\Sigma^{-1} = R = \begin{pmatrix} r_1 & r \\ r & r_2 \end{pmatrix}$, it can be shown using (3.6) that the corresponding weights are

$$\tfrac{1}{3} + \tfrac{1}{2n}(\tfrac{r_1}{2} + r)$$

and

$$\tfrac{2}{3} - \tfrac{1}{2n}(\tfrac{r_1}{2} + r).$$

The actual condition on n is that these two quantities lie in [0,1].

We conclude this section by showing that the Bayes design with respect to L has other remarkable Bayesian robustness properties. The next two results are related to hypothesis testing and set estimation as opposed to point estimation which was stressed thus far.

Theorem 3.4. Consider the setup of Theorem 3.1.

(a) For any vector $\utilde{c}$, and a fixed design M, let I be the set of smallest Lebesgue measure such that $\inf_{L \le \pi \le kL} P(\utilde{c}'\utilde{\theta} \varepsilon I | \utilde{y}) \ge 1 - \alpha$, where $1 > \alpha > 0$ is a fixed number. Then the design minimizing the Lebesgue measure of I is the Bayes design with respect to L for the point estimation problem with squared error loss, i.e., the design that minimizes $\utilde{c}'(M + \Sigma^{-1})^{-1}\utilde{c}$.

(b) For a fixed design M, let S be the set of smallest Lebesgue measure such that $\inf_{L \le \pi \le kL} P(\utilde{\theta} \varepsilon S | \utilde{y}) \ge 1 - \alpha$, where $1 > \alpha > 0$ is a fixed number. Then the design minimizing the Lebesgue measure of S is the design that is Bayes D–optimal with respect to L, i.e., the design that minimizes $|M + \Sigma^{-1}|^{-1}$.

Remark 8. The important points of the above theorem are that it shows one more robustness property of the design that is Bayes with respect to L, and more importantly, it relates the point estimation and the set estimation problems and demonstrates that a design which is optimal in one problem will be optimal in the other problem too. This is reassuring.

Remark 9. The problem of finding the smallest volume confidence set with a minimum posterior probability of $1 - \alpha$ for a family of plausible priors has received attention from several statisticians, including LeCam (1986).

Proof of Theorem 3.4. The proofs of parts (a) and (b) are similar; so we will sketch the proof of only part (b).

It is proved in DasGupta and Studden (1988b) that (the) set S exists and is simply a Bayes confidence set for the prior L, i.e., for a suitable $\gamma < \alpha$, S satisfies $P_L(\utilde{\theta} \varepsilon S | \utilde{y}) = 1 - \gamma$. Since the posterior distribution of $\utilde{\theta}$ under the prior L is $N((M + \Sigma^{-1})^{-1}X'(\utilde{y} - X\utilde{\mu}), (M + \Sigma^{-1})^{-1})$, it follows that S is the p dimensional ellipsoid

$$S = \left\{\utilde{\theta} \colon (\utilde{\theta} - \utilde{\nu})'\Lambda^{-1}(\utilde{\theta} - \utilde{\nu}) \le X^2_{1-\gamma}(p)\right\}, \tag{3.7}$$

where $\Lambda = (M+\Sigma^{-1})^{-1}$, $\utilde{\nu} = \Lambda X'(\utilde{y} - X\utilde{\mu})$, and $X^2_{1-\gamma}(p)$ is the $100(1-\gamma)$th percentile of the X^2 distribution with p degrees of freedom. Since the Lebesgue measure of S is proportional to $|M+\Sigma^{-1}|^{-\frac{1}{2}}$, the result follows.

Example 11. Consider the quadratic regression model $E(y) = \theta_0 + \theta_1 x + \theta_2 x^2$, and suppose $-1 \leq x \leq 1$; also let L be the $N(\utilde{\mu}, \sigma^2\Sigma)$ prior where $\utilde{\mu}$ is arbitrary but fixed and $\Sigma^{-1} = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3)$. Then standard monotonicity and convexity arguments and calculus give that the optimum design of part (b) is of the form

$$M = n\begin{pmatrix} 1 & 0 & c \\ 0 & c & 0 \\ c & 0 & c \end{pmatrix},$$

$$\text{where } c = \frac{\frac{\lambda_2-\lambda_1}{n} + 1 + \sqrt{\left(\frac{\lambda_2-\lambda_1}{n}+1\right)^2 + 3\left(1+\frac{\lambda_1}{n}\right)\left(\frac{\lambda_2+\lambda_3}{n}\right)}}{3} \tag{3.8}$$

Of course this amounts to sampling at 0 and ± 1, where the proportion of observations at each of ± 1 is $\frac{c}{2}$. For example, if the prior variances of $\theta_0, \theta_1, \theta_2$ under L are 3, 5, and 1, and if $n = 9$, then c is approximately .72. Notice that the optimal design converges to the classical D–optimal design as $n \to \infty$.

We will now state and prove the final result of this article. The purpose of this result will be to show that very surprisingly, testing and point estimation problems lead to the same optimum robust design and that design is precisely the Bayes design with respect to the prior L. A precise statement of this result is given below.

Theorem 3.5. Consider the setup of Theorem 3.1. Suppose we want to test the hypothesis that for a fixed vector $\utilde{c}$, $\utilde{c}'\utilde{\theta}$ is smaller than or equal to its prior expected value, i.e., $H_0\colon \utilde{c}'\utilde{\theta} \leq \utilde{c}'\utilde{\mu}$. Consider this as a decision problem with a zero–one loss $L(H_i, a_j) = \delta_{ij}$, i, $j = 0, 1$, where a_j denotes the action "accept H_j" and $\delta_{ij} = 1$ if $i \neq j$ and 0 if $i = j$. Then the designs that minimize (a) the Bayes risk (of the Bayes test) with respect to L, (b) the range of the posterior probabilities of H_0, i.e., $\sup\limits_{L \leq \pi \leq kL} P(H_0|\utilde{y}) - \inf\limits_{L \leq \pi \leq kL} P(H_0|\utilde{y})$, and (c) the Bayes risk with respect to L for the problem of estimating $\utilde{c}'\utilde{\theta}$ under an ordinary squared error loss, are identical; hence the optimum design for all three problems is the design that minimizes $\utilde{c}'(M+\Sigma^{-1})^{-1}\utilde{c}$.

Remark 10. The strength of this result is in the fact that in order to get the smallest possible range of the posterior probabilities of H_0, one merely needs to construct the design that is Bayes with respect to L for the testing problem. However, what we consider to be extremely surprising is that the optimum designs for the testing and the point estimation problems coincide. Thus, an experimenter who is simultaneously interested in conducting a variety of statistical analyses can go ahead and use the same design, a very reassuring situation. The proof of Theorem 3.5 needs the following Lemma.

Lemma 3.6. Let $Z \sim N(0,\tau^2)$ and let $g(\Phi(Z))$ be a symmetric unimodal function of Z with mode at 0. Then $E[g(\Phi(Z))]$ is decreasing in τ.

Proof. See appendix.

Proof of Theorem 3.5: Assume without loss of generality that $\underset{\sim}{\mu} = \underset{\sim}{0}$. Since the loss is zero–one, for a fixed design M, the Bayes test under L takes action a_0 if and only if $P_L(H_0|\underset{\sim}{y}) \geq P_L(H_1|\underset{\sim}{y})$. Consequently, the posterior risk equals $g_1(p) = \min(p,(1-p))$ where $p = P_L(H_0|\underset{\sim}{y})$ (note p will depend on the design M). Notice $g_1(p)$ is symmetric about $p = \frac{1}{2}$ and also unimodal with mode at $\frac{1}{2}$. Now, since the posterior distribution of $\underset{\sim}{c}'\underset{\sim}{\theta}$ under L is $N(\underset{\sim}{c}'\underset{\sim}{\nu}, \underset{\sim}{c}'\Lambda\underset{\sim}{c})$ where $\underset{\sim}{\nu}$ and Λ are as in the proof of Theorem 3.4 (with $\underset{\sim}{\mu} = \underset{\sim}{0}$), it follows that $p = \Phi(-\frac{\underset{\sim}{c}'\underset{\sim}{\nu}}{\sqrt{\underset{\sim}{c}'\Lambda\underset{\sim}{c}}})$.

$$\therefore \text{ the Bayes risk of the Bayes test under} L$$

$$= E_{M_L(\underset{\sim}{y})}[g_1(p)]$$

(where $M_L(\underset{\sim}{y})$ denotes the marginal distribution of $\underset{\sim}{y}$ under the prior L)

$$= E_{M_L(t)}[g_1(\Phi(-t))], \tag{3.9}$$

where $t = \frac{\underset{\sim}{c}'\underset{\sim}{\nu}}{\sqrt{\underset{\sim}{c}'\Lambda\underset{\sim}{c}}}$ and $M_L(t)$ denotes the marginal distribution of t under the prior L. Trivially, $M_L(t)$ is the $N(0,\tau^2)$ distribution,

$$\begin{aligned} \text{where } \tau^2 &= \frac{\underset{\sim}{c}'\Sigma M(M+\Sigma^{-1})^{-1}\underset{\sim}{c}}{\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{c}} \\ &= \frac{\underset{\sim}{c}'\Sigma\underset{\sim}{c}}{\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{c}} - 1. \end{aligned} \tag{3.10}$$

Now notice that $g_1(\Phi)$ is symmetric and unimodal in t and therefore by Lemma 3.6 and (3.10), (3.9) is increasing in $\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{c}$. Consequently, the optimum design for part (a) of Theorem 3.5 is the design that minimizes $\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{c}$.

To derive the optimum design for part (b) of the theorem, let A denote the set

$$A = \{\underset{\sim}{\theta} : \underset{\sim}{c}'\underset{\sim}{\theta} \leq 0\} \tag{3.11}$$

$$\begin{aligned} \therefore \sup_{L \leq \pi \leq kL} & P(H_0|\underset{\sim}{y}) \\ = \sup_{L \leq \pi \leq kL} & \frac{\int_A d\pi(\underset{\sim}{\theta}|\underset{\sim}{y})}{\int_{\mathbb{R}^p} d\pi(\underset{\sim}{\theta}|\underset{\sim}{y})} \end{aligned} \tag{3.12}$$

where $\pi(\underset{\sim}{\theta}|\underset{\sim}{y})$ denotes the posterior distribution of $\underset{\sim}{\theta}$ given $\underset{\sim}{y}$ resulting from a generic prior π, where $L \leq \pi \leq kL$.

It is easy to see that the ratio $\int\limits_A d\pi(\underset{\sim}{\theta}|\underset{\sim}{y}) / \int\limits_{\mathbb{R}^p} d\pi(\underset{\sim}{\theta}/\underset{\sim}{y})$ is maximized by the prior

$$\begin{aligned} \pi(\underset{\sim}{\theta}) &= kL(\underset{\sim}{\theta}) \text{ if } \underset{\sim}{\theta}\varepsilon A \\ &= L(\underset{\sim}{\theta}) \text{ if } \underset{\sim}{\theta} \not\varepsilon A \end{aligned} \tag{3.13}$$

(see DasGupta and Studden (1988a) and DeRobertis (1978)).

$$\begin{aligned} \therefore \sup_{L \leq \pi \leq kL} & P(H_0|\underset{\sim}{y}) \\ &= \frac{kp}{kp+1-p} \\ &= \frac{kp}{1+(k-1)p}. \end{aligned} \tag{3.14}$$

$$\begin{aligned} \text{Similarly, } \inf_{L \leq \pi \leq kL} & P(H_0|\underset{\sim}{y}) \\ &= \frac{p}{p+k(1-p)} \\ &= \frac{p}{k-(k-1)p} \end{aligned} \tag{3.15}$$

($P(H_0|\underset{\sim}{y})$ is minimized by using the prior

$$\begin{aligned} \pi(\underset{\sim}{\theta}) &= L(\underset{\sim}{\theta}) \text{ if } \underset{\sim}{\theta}\varepsilon A \\ &= kL(\underset{\sim}{\theta}) \text{ if } \underset{\sim}{\theta}\varepsilon A^c.) \end{aligned} \tag{3.16}$$

$$\begin{aligned}
\therefore\ & \sup_{L\le\pi\le kL} P(H_0|\underset{\sim}{y}) - \inf_{L\le\pi\le kL} P(H_0|\underset{\sim}{y}) \\
&= \frac{(k^2-1)p(1-p)}{(1+(k-1)p)(k-(k-1)p)} \qquad (3.17)\\
&= g_2(p)(\text{say}).
\end{aligned}$$

$g_2(p)$ is easily seen to be symmetric about $p = \frac{1}{2}$ and unimodal with mode at $p = \frac{1}{2}$. Thus the expected range of the posterior probability of H_0 (under the marginal distribution of $\underset{\sim}{y}$ induced by the prior L) is increasing in $\underset{\sim}{c}'(M+\Sigma^{-1})^{-1}\underset{\sim}{c}$ by a repetition of the argument used to prove part (a) of the Theorem. The theorem is now proved.

4. Concluding remarks, other models, generalizations. In the present article we have taken a novel approach of designing an experiment when we want to use the available prior information but also want to guard as much as possible against possible misspecification of prior information. Our results include several new convexity results and especially encouraging are the findings in section 3 that the user who wants to estimate and test at the same time can use the same optimum design.

Much more has to be done. Other ways to model prior information have to be considered; Huber (1973), Berger and Berliner (1986), O'Hagan and Berger (1988) discuss useful ways to model prior information. The case of an unknown error variance was not considered in this article to keep the setup simple. However, most results of this paper are also valid when the error variance σ^2 is unknown and an appropriate inverse gamma prior is used for σ^2. The practically useful cases of heteroscedastic and/or correlated errors will be considered elsewhere.

5. Appendix

Proof of Lemma 2.3: Let $R = \Gamma' D_1 \Gamma$ where Γ is orthogonal and D_1 is diagonal. Then, $trRST = trD_1S_1T_1$ where $S_1 = \Gamma S \Gamma'$ and $T_1 = \Gamma T \Gamma'$. Note $S_1T_1 = T_1S_1$ and both S_1, T_1 are symmetric. Therefore, there exists an orthogonal matrix L and diagonal matrices D_2 and D_3 such that $S_1 = LD_2L'$ and $T_1 = LD_3L'$. Note, D_1, D_2, D_3 are *nnd* because

R, S_1, T_1 are

$$\begin{aligned}&\therefore trRST\\&= trD_1S_1T_1\\&= trL'D_1LD_2D_3\\&\geq 0,\end{aligned}$$

since $L'D_1L$ and D_2D_3 are *nnd*.

Proof of Lemma 3.6: Since $g(\Phi(Z))$ is symmetric,

$$\begin{aligned}&E[g(\Phi(z))]\\&= 2\int_0^\infty g(\Phi(z))\frac{e^{-\frac{z^2}{2\tau^2}}}{\sqrt{2\pi\tau}}dz.\end{aligned} \tag{5.1}$$

Note that for $z > 0$, $g(\Phi(z))$ is decreasing in z; since $\frac{2e^{-\frac{z^2}{2\tau^2}}}{\sqrt{2\pi\tau}}I_{\{z>0\}}$ is MLR in z, the result follows immediately.

References

Bandemer, H. (1977). Theorie und Anwendung der Optimalen Versuchsplanung I and II. Akademie–Verlag, Berlin.

Berger, James (1984). The robust Bayesian viewpoint. Robustness in Bayesian Statistics. J. Kadane (ed.), North–Holland, Amsterdam.

Berger, James (1987). Robust Bayesian analysis: sensitivity to the prior. Tech. Report **#87–10**, Dept. of Statistics, Purdue University.

Berger, James and Berliner, L. M. (1986). Robust Bayes and empirical Bayes analysis with ε–contaminated priors. *Ann. Statist.* **14**, 461–486.

Brooks, R. J. (1972). A decision theory approach to optimal regression designs. *Biometrika* **59**, 563–571.

Brooks, R. J. (1974). On the choice of an experiment for prediction in linear regression. *Biometrika* **61** 303–311.

Chaloner, Kathryn (1984). Optimal Bayesian experimental designs for linear models. *Ann. Statist.* **12**, 283–300.

Cheng, C–S. (1987). An application of the Kiefer–Wolfowitz equivalence theorem to a problem in Hadamard transform optics, *Ann. Statist.* **15**, 1593–1603.

Chernoff, H. (1972). *Sequential Analysis and Optimal Design*, SIAM, Philadelphia.

DasGupta, Anirban and Studden, W. J. (1988a). Robust Bayesian analysis and optimal experimental designs in normal linear models with many parameters–I. Tech. Report, Dept. of Statistics, Purdue University.

DasGupta, Anirban and Studden, W. J. (1988b). Frequentist behavior of smallest volume robust Bayes confidence sets. Tech. Report, Dept. of Statistics, Purdue University.

DasGupta, Anirban and Studden, W. J. (1988c). Variations in posterior measures for priors in a band: effect of additional restrictions. Tech. Report, Dept. of Statistics, Purdue University.

DasGupta, Anirban and Studden, W. J. (1988d). Frequentist behavior of robust Bayes procedures: new applications of the Wald–Lehmann minimaxity theory. Tech. Report, Dept. of Statistics, Purdue University.

Dempster, A. P. (1975). A subjectivist look at robustness. *Bull. Int. Statist. Inst.* **46**, 349–374.

DeRobertis, L. (1978). The use of partial prior knowledge in Bayesian inference. Ph.D. Thesis, Yale University, New Haven.

DeRobertis, L. and Hartigan, J. A. (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **1**, 235–244.

Duncan, G. and DeGroot, M. H. (1976). A mean squared error approach to optimal design theory. *Proceedings of the 1976 conference on information: sciences and systems.* The Johns Hopkins University, 217–221.

Duncan, G. and Lambert, D. (1981). Bayesian learning based on partial prior information. Tech. Report, Department of Statistics, Carnegie–Mellon University.

Elfving, G. (1952). Optimum allocation in linear regression theory. *Ann. Math. Statist.* **23**, 255–262.

Fedorov, V. V. (1972). *Theory of Optimal Experiments.* Trans. and ed. W. J. Studden and E. M. Klimko. Academic, New York.

Gaffke, N. (1985). Directional derivatives of optimality criteria at singular matrices in convex design theory, *Statistics* **16**, 373–388.

Good, I. J. and Crook, J. F. (1987). The robustness and sensitivity of the mixed–Dirichlet Bayesian test for "independence" in contingency tables. *Ann. Statist.* **15**, 670–693.

Hoel, Paul G. (1966). A simple solution for optimal Chebyshev regression extrapolation. *Ann. Math. Statist.* **37**, 720–725.

Hoel, P. G. and Levine, A. (1964). Optimal spacing and weighting in polynomial prediction. *Ann. Math. Statist.* **35**, 1553–1563.

Huber, P. J. (1973). The use of Choquet capacities in Statistics. *Proc. 39th Session ISI*, Vol. **45**, 181–188.

Kadane, J. and Chuang, D. T. (1978). Stable decision problems. *Ann. Statist.* **6**, 1095–1110.

Karlin, S. and Studden, W. J. (1966). *Tchebycheff Systems: with Applications in Analysis and Statistics.* Wiley, New York.

Kiefer, J. (1959). Optimal experimental designs, (with discussion). *J. Roy. Statist. Soc. B. 21*, 272–319.

Kiefer, J. (1961). Optimum designs in regression problems II. *Ann. Math. Statist.* **32**, 298–325.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Statist.* **2**, 849–879.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *Ann. Math. Statist.* **30**, 271–294.

Kiefer, J. and Wolfowitz, J. (1965). On a theorem of Hoel and Levine on extrapolation. *Ann. Math. Statist.* **36**, 1627–1655.

Leamer, E. E. (1978). Specification searches: ad hoc inference with nonexperimental data. John Wiley, New York.

Leamer, E. E. (1982). Sets of posterior means with bounded variance prior. *Econometrica* **50**, 725–736.

LeCam, L. (1986). Discussion of "On the consistency of Bayes estimates", *Ann. Statist.* **14**, 1–67.

Lee, C. M.–S. (1988). Constrained optimal designs, *JSPI* **18**, 377–389.

Lindley, D. V. (1968). The choice of variables in multiple regression. *J. Roy. Statist. Soc. B* **32**, 31–53.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. (Ser. B)* **34**, 1–41.

O'Hagan, A. and Berger, J. (1988). Ranges of posterior probabilities for quasiunimodal priors with specified quantiles. *Jour. Amer. Statist. Assoc.* **83**, 503–508.

Pilz, J. (1979a). Konstruktion von optimalen diskreten Versuchsplänen für eine Bayes–Schätzung im linearen Regressionsmodell. *Freiberger Forschungshefte* **117**, 123–152.

Pilz, J. (1981a). Robust Bayes and minimax–Bayes estimation and design in linear regression. *Math. Operationsforsch. Stat., Ser. Statistics* **12**, 163–177.

Polachek, W. (1985). Sensitivity analysis for general and hierarchical linear regression models. Bayesian Inference and Decision techniques with Applications, Eds. P. K. Goel and A. Zellner, North–Holland, Amsterdam.

Pukelsheim, F. and Titterington, D. M. (1983). General differential and Lagrangian theory for optimal experimental design. *Ann. Statist.* **11**, 1060–1068.

Sacks, J. and Ylvisaker, D. (1968). Designs for regression problems with correlated errors: many parameters. *Ann. Math. Statist.* **39**, 49–69.

Schoenberg, I. J. (1959). On the maxima of certain Hankel determinants and the zeros of the classical orthogonal polynomials, *Indag. Math.* **21**, 282–290.

Silvey, S. D. (1980). *Optimal Design.* Chapman and Hall, London and New York.

Studden, W. J. (1968). Optimum design on Tchebycheff points. *Ann. Math. Statist.* **39**, 1435–1447.

Verdinelli, I. (1982). Computing Bayes *D*– and *A*–optimal block designs for a two–way model. Unpublished manuscript.

West, S. (1979). Upper and lower probability inferences for the logistic function. *Ann. Statist.* **7**, 400–413.

Whittle, P. (1973). Some general points in the theory of optimal experimental design. *J. Roy. Statist. Soc. B* **35**, 123–130.

Wynn, H. P. (1972). Results in the theory and construction of D–optimum experimental designs. *J. Roy. Statist. Soc. B* **34**, 133–147.

Fig. 2.1

Plot of $\eta(\varepsilon)$ vs. ε

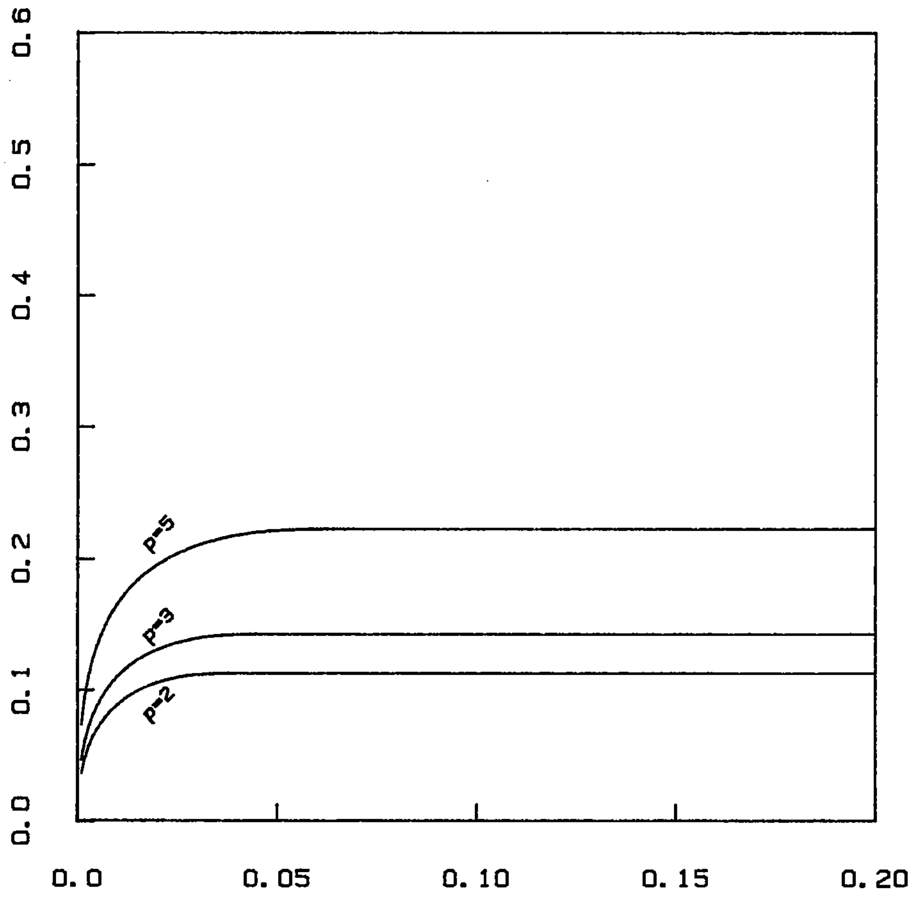
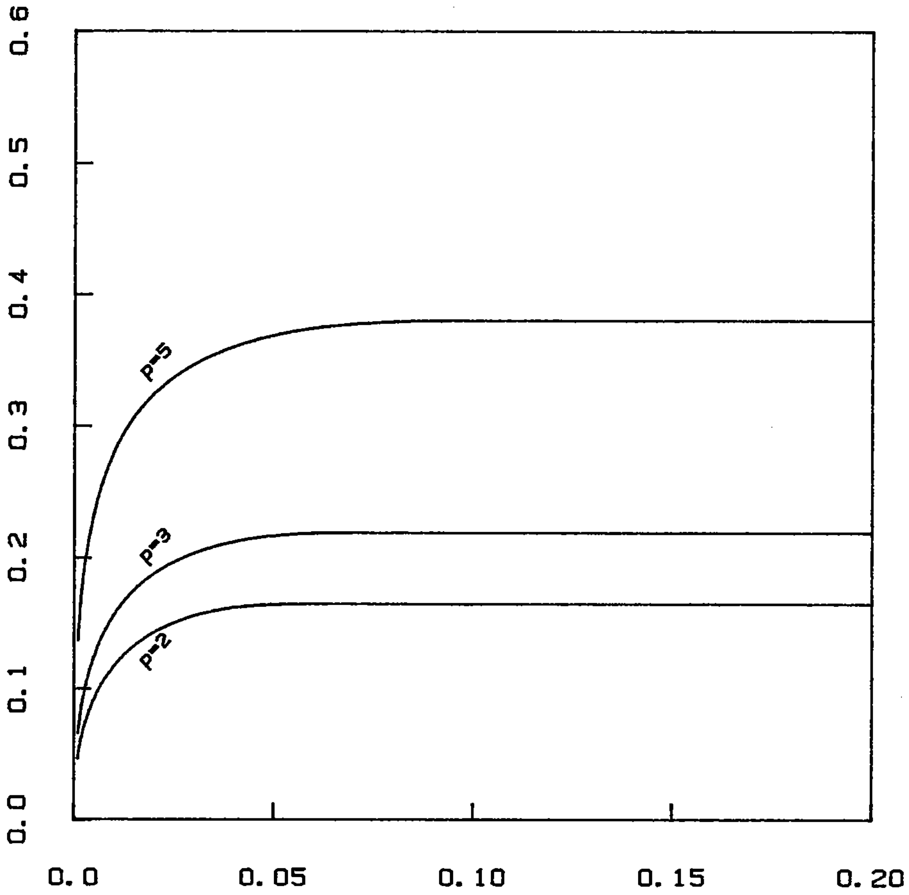


Fig. 2.2

Plot of $\eta(\varepsilon)$ vs. ε


0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.05
0.10
0.15
0.20
P=5
P=3
P=2